AD_____

Grant Number DAMD17-94-J-4406

TITLE: Statistical Genetics Methods for Localizing Multiple Breast Cancer Genes

PRINCIPAL INVESTIGATOR: Jurg Ott, Ph.D.

CONTRACTING ORGANIZATION: Columbia University
New York, New York 10023

REPORT DATE: September 1996

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 1996	Annual (1 Sep 95 - 31 Aug 96)

4. TITLE AND SUBTITLE
Statistical Genetics Methods for Localizing Multiple Breast Cancer Genes

5. FUNDING NUMBERS
DAMD17-94-J-4406

6. AUTHOR(S)
Jurg Ott, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Columbia University
New York, New York 10032

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, Maryland 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200*

The first breast cancer gene, BRCA1, was localized in 1990. Now, a second breast cancer gene is known on chromosome 13 yet many breast cancer families show absence of linkage to either of these genes. Thus, investigators are searching to find additional genes responsible for familial breast cancer. These genes are expected to be associated particularly with late onset breast cancer while BRCA1 occurs primarily in early onset disease. One of the problems with linkage analysis of late onset diseases is that parents may be unavailable. Thus, in affected sib pair analysis (a widely used nonparametric linkage analysis technique), there is no way that errors can be detected through mendelian inconsistencies. In other words, the purported siblings may not be sibs but, for example, half-sibs or unrelated individuals. A method was developed and implemented in a computer program to screen for non-sibs by statistical means. The results show that its application greatly increases power of affected sib pair linkage analysis.

14. SUBJECT TERMS
Breast Cancer, errors, mendelian inconsistency, sample swaps, affected sib pairs, missing parents

15. NUMBER OF PAGES
11

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

)pinions, interpretations, conclusions and recommendations are :hose of the author and are not necessarily endorsed by the US \rmy.

____ Where copyrighted material is quoted, permission has been)btained to use such material.

____ Where material from documents designated for limited listribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Signature 9/25/96

PI - Signature Date

(none of the above apply to this project)

TABLE OF CONTENTS

INTRODUCTION

The work reported below addresses Task 3, *Approximating the Maximum Lod Score*, of the original grant application. As originally proposed, a method to *approximate* lod scores (rather than to calculate them exactly) was developed on the basis of computer simulation in a specific pedigree. It turned out, however, that the number of replicates required for reasonable accuracy was so large that the method was not any faster than exact calculation of lod scores. Therefore, the originally proposed method was modified in favor of a new approach to optimize linkage analysis. This approach also involves approximation in the sense that an estimation procedure was developed to screen for sib pairs with errors, which are then removed from analysis. Computer simulation of the method demonstrates that its application leads to increased power in linkage analysis. The rationale for the new approach is presented in the following paragraph and the method itself is described in the *Body* section.

Currently, two breast cancer genes, BRCA1 and BRCA2, are known, yet many breast cancer families show absence of linkage to either of these genes. Thus, investigators are searching for additional genes responsible for familial breast cancer. These genes are expected to be associated with late onset breast cancer while BRCA1 occurs primarily in early onset disease. One of the problems with linkage analysis of late onset disease is that parents may be unavailable. Thus, in affected sib pair analysis (a widely used nonparametric linkage analysis technique), there is no way that errors can be detected through mendelian inconsistencies: The two sibs share at most four different alleles, which is the same number as there are alleles among two parents. Laboratory errors (sample swaps, allele misreading, etc.), unrecognized adoption, and other errors are generally recognized through the occurrence of mendelian inconsistencies but this is not possible for two affected sibs with their parents missing. Such errors typically occur with frequencies of around 1% and have the consequence that the purported siblings may not be sibs but, for example, half-sibs or unrelated individuals. In linkage analysis, occurrence of errors greatly reduces informativeness (Terwilliger et al. 1990; Buetow 1991). To offset this potential loss of information, a method was developed and implemented in a computer program to screen for non-sibs by statistical means and to remove them from the analysis. As shown below, application of this method greatly increases power of affected sib pair linkage analysis.

BODY

As a consequence of the mendelian laws of inheritance, the genotypes of two relatives are similar to each other. Consequently, by establishing whether or not the genotypes of two individuals are correlated and to what degree, it should be possible to determine whether these individuals are related or not, and perhaps what the degree of relationship is. On the basis of the genotypes for a set of unlinked marker loci, Thompson (1975, 1991) developed appropriate statistical theory to estimate the relationship between two individuals. To apply her approach to this project, her theory was extended to allow for unlinked *and linked* markers.

In addition, Bayesian methods are applied that incorporate the typically known prior error probabilities. For details on the theory of relationship estimation, see the manuscript provided in the appendix (Goring and Ott 1996).

A simple outline of the main steps of the new method is as follows: Based on the genotypes at all marker loci available in an affected sib pair study, for each stated sib pair the posterior probability is computed that the two individuals are siblings. When this probability falls below a certain threshold then the pair of individuals is discarded from the study (because they are assumed not to be siblings). The remaining stated sib pairs are then most likely pairs of true siblings and are analyzed in one of the usual affected sib pair approaches. The threshold was chosen in the basis of a decision rule that maximizes power if in fact linkage of a recessive trait to a fully informative marker exists. Computer simulation shows that this decision rule is conservative, that is, very few true siblings tend to be discarded from a study (see manuscript in appendix).

Even though the new method reduces the number of "sib" pairs available for analysis, it results in a dramatic increase of power because the sib pairs remaining in the analysis after application of the Bayesian relationship estimation are with high probability real sibs. For example, consider the following case (case *a* in table 5 of manuscript in appendix): 400 sib pairs without parents are available for study. They tend to consist of 98% true sibs, 1% half-sibs, and 1% pairs of unrelated individuals. 100 marker loci (70% heterozygosity each) are typed on each sib pair. The disease was taken to be a complex trait (multifactorial threshold trait) with population prevalence of 5% and heritability (on the liability scale) of 50%. The recombination fraction between the major disease locus and a marker locus was assumed to be 1%. The result of an affected sib pair analysis is considered significant when the empirical significance level is at most 0.0001, which approximately corresponds to a maximum lod score of 3. Under these conditions, power to detect linkage is 0.39 when all stated sib pairs are used. It increases to 0.51 when the new method is applied prior to the affected sib pair analysis. This value of 0.51 is also the power when the known non-sibs are removed prior to affected sib pair analysis, that is, the new method removes non-sibs with high confidence and has little tendency to remove true sibs.

CONCLUSIONS

The new method presented above represents an efficient way of improving nonparametric linkage analysis when parents are unavailable for study. Thus, it is particularly suitable for late onset disease.

Genetic Epidemiology 12:859–862 (1995)

Variability of Genotype-specific Penetrance Probabilities in the Calculation of Risk Support Intervals

Suzanne M. Leal and Jurg Ott

Department of Otolaryngology (S.M.L.), University of Tübingen, Germany; Department of Psychiatry (J.O.), Columbia University, New York, New York

Previously, a maximum likelihood method was described to construct a support interval for the risk. This method is extended to incorporate genotype specific penetrance probabilities in the calculation of a risk support interval. As an empirical example, the support interval for the risk is calculated for a member of a published breast-ovarian cancer kindred. © 1995 Wiley-Liss, Inc.

Key words: genetic counseling, support interval, genetic risk, phenotype risk

INTRODUCTION

In a Mendelian trait, the genetic risk is the conditional probability that an individual has the genetic susceptible genotype given both phenotype and genotype information for all available pedigree members. Genetic risks may be based on the pedigree likelihood [Elston and Stewart, 1971]. In addition to such genotype risks, a phenotype risk may be defined as the conditional probability of developing the trait. With incomplete penetrance and absence of phenocopies, the phenotype risk is smaller than the corresponding genotype risk. Generally, however, phenocopies as well as genetic cases contribute to the phenotype risk.

The precision of risk estimates is dependent on the accuracy of the parameters used in their evaluation. Usually risks are computed under the assumption that genetic parameters are known without error. Uncertainty in the accuracy of parameter estimates renders uncertainty in the risk. Therefore, to evaluate the accuracy of a risk it is critical to calculate either a confidence or support interval for the risk, [Smith, 1971; Lange, 1986; Rogatko, 1988; Weeks and Ott, 1989; Suthers and Wilson, 1990; Ott, 1991; Leal and Ott, 1994].

Address reprint requests to Dr. Jurg Ott, Department of Psychiatry, Columbia University, 722 West 168th Street, Unit 58, New York, NY 10032.

Previously, we described a method to construct support intervals (SIs) for genetic risks working in a maximum likelihood framework [Leal and Ott, 1994]. Briefly, the method allows for parameters to vary in their support intervals. For each combination of parameter values so obtained, a risk is calculated whose associated log likelihood is equal to the log likelihood at the given parameter values. All those risk values with a log likelihood within m units of the maximum log likelihood form the risk support interval. Here, this method is expanded to allow for variability of genotype-specific penetrances when the age at disease onset is normally distributed. As an empirical example, the SIs for the phenotype and genotype risk will be calculated below for a member of a breast-ovarian cancer kindred using two markers (D17S250 and D17S588) which are linked to the BRCA1 locus.

METHODS

In genetic counseling situations, one generally works with a single pedigree. Usually, parameter estimates must be obtained from previously published results. A maximum likelihood method to construct an SI for the risk under these circumstances was previously described [Leal and Ott, 1994]. In principle, we rely on published support intervals. If these are unavailable, we construct them by one of several methods using information in published reports.

Below genotype-specific penetrance probabilities are incorporated in the calculation of SIs for the genotype and phenotype risk. Approximate m-unit SIs are constructed around the mean age of disease onset, μ, and lifetime penetrances, λ, each for disease gene carriers and noncarriers. The calculation of maximum and joint log likelihoods for all parameters is carried out as previously described [Leal and Ott, 1994], except that here, the estimates, μ, for age at disease onset are taken to follow a normal distribution while all other parameter estimates are binomially distributed.

The penetrance probabilities are the genotype-specific cumulative risk for unaffected and affected individuals when age of onset is unknown, and genotype-specific density for affected individuals when age of onset is known.

At this point, each parameter is varied within its SI. When the joint log likelihood for a set of parameter values falls within m units of maximum log likelihood, the genotype-specific penetrance probabilities are calculated for each liability class and the risk is calculated with the aid of MLINK [Lathrop and Lalouel, 1984]. The phenotype risk is also computed using a specific cumulative penetrance liability class. The highest and lowest (genotype and phenotype) risks so obtained are taken to be the endpoints of the (genotype and phenotype) risk SI.

EXAMPLE

As an empirical example, 2-unit SIs for the phenotype and genotype risk were calculated for individual 405, an unaffected 52-year-old female who is a member of the breast-ovarian cancer kindred CRC101 [Smith et al., 1993], given her current age.

The following parameter values were used: male map distance between the two flanking markers D17S250 and S17S588 = 9.8 cM Haldane; female-to-male map distance ratio = 2.0; the disease locus map position, x = 4.5 cM centromeric to D17S588, the SI (maximum lod score-1) for this map position ranges 2.9 cM proximal from D17S588 to

3.6 cM distal to D17S250 (length of SI 3.28 cM); the proportion of linked families, α = 1.0 with an SI of (0.71,1.0) [Easton et al., 1993; Bishop, personal communication]; the age of onset was assumed to be normally distributed with a mean age of onset for breast cancer gene carriers, $\mu_{AA} = \mu_{Aa} = 55.435$ (standard error = 1.742) and noncarriers, μ_{aa} = 68.990 (standard error = 1.532), the standard deviation, $\sigma_{AA} = \sigma_{Aa} = \sigma_{aa} = 15.387$; the lifetime penetrance for breast cancer gene carriers, $\lambda_{AA} = \lambda_{Aa} = 0.928$ (standard error = 0.163) and noncarriers, $\lambda_{aa} = 0.1$ (standard error = 0.009); and the disease gene frequency, q = 0.0033 (standard error = 0.0012) [Claus et al., 1991].

As previously described by Easton et al. [1993], affected and unaffected individuals were assigned to disease and age specific liability classes with all males and ovarian cancer cases assigned to the youngest age class. A total of 14 age specific liability classes were formed, 7 for unaffected and 7 for affected individuals (ages < 30, 30-39, 40-49, 50-59, 60-69, 70-79, and >80) using the middle of each age class (25, 35, etc).

The parameters, x, α, q, μ_{AA}, μ_{aa}, λ_{AA}, and λ_{aa} were each varied while all other parameters were held fixed. Whenever the joint log likelihood fell within 2 units of the maximum log likelihood, the risk for the counselee was calculated under heterogeneity as a weighted average of the risk under linkage and no linkage [Weeks and Ott, 1989]. The phenotype risk (that individual 405 will be affected within her lifetime [< 80 years of age]) was calculated using the cumulative penetrance age class 70-79.

The highest and lowest (genotype and phenotype) risks were taken to be the upper and lower bounds for the (genotype and phenotype) SI. Point (phenotype and genotype) risk estimates were calculated under homogeneity, α = 1.0 using the point estimates for all parameters.

RESULTS

The SI for the genotype risk that individual 405 carries the BRCA1 susceptibility allele is (0.0%, 14.5%) and the SI for the phenotype risk is (5.9%, 19.4%). The point estimates for the genotype and phenotype risks are 0.021 and 0.084, respectively.

DISCUSSION

The calculation of SIs enables genetic counselors to determine the reliability of risk estimates. An SI for the risk can help to determine the accuracy of the risk estimate, where a wide SI reflects an inaccurate point estimate.

In the method described, only "statistical" or sampling variability of the estimate is allowed for in the calculation of SIs for the risk. Phenotypic errors, marker map position errors, and systematic bias will also affect the accuracy of the risk; however, these errors are not taken into account in this method.

For this example, the SI for the genetic risk is wide (0.0%,14.5%). However, the upper bound of the SI for the risk is lower than the counselees genotype risk (R = 40.5%) if no marker data were available. The SI for the phenotype risk (5.9%, 19.4%) reflects the variability of the genotype-specific cumulative penetrances for breast cancer gene and nongene carriers. It should be noted that often SIs may be wider than one would expect, especially under linkage heterogeneity [Leal and Ott, 1994].

The computer program, RISKSI, which implements the procedures outlined above, and an auxiliary program, RISKPREP, which aids users in the creation of the necessary data file are available on the anonymous ftp site linkage.cpmc.columbia.edu.

ACKNOWLEDGMENTS

This material is based upon work supported by US Army Medical Research Acquisition Activity under award #DAMD17-94-J-4406. Any opinions, findings and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the US Army Medical Research Acquisition Activity. We also acknowledge support by the Deutsche Forschungsgemeinschaft (Klifo Hörforschung; Zel 49/6-1).

REFERENCES

Claus E, Risch N, Thompson WD (1991): Genetic analysis of breast cancer in the cancer and steroid hormone study. Am J Hum Genet 48:232-242.

Easton DF, Bishop DT, Ford D, Crockford GP and the Breast Cancer Linkage Consortium (1993): Genetic linkage analysis in familial breast and ovarian cancer: Results from 214 families. Am J Hum Genet 52:678-701.

Elston RC, Stewart J (1971): A general model for the analysis of pedigree data. Hum Hered 21:523-542.

Lange K (1986): Approximate confidence intervals for risk prediction in genetic counseling. Am J Hum Genet 38:681-687.

Lathrop GM, Lalouel JM (1984): Easy calculations of lod scores and genetic risks on small computers. Am J Hum Genet 36:460-465.

Leal SM, Ott J (1994): A likelihood approach to calculating risk support intervals. Am J Hum Genet 54:913-917.

Ott J (1991): "Analysis of Human Genetic Linkage." Baltimore: Johns Hopkins University Press.

Rogatko A (1988): Evaluating the uncertainty of risk predication in genetic counseling: A Bayesian approach. Am J Med Genet 31:513-519.

Smith C (1971): Recurrence risk for multifactorial inheritance. Am J Hum Genet 23:578-588.

Smith SA, Easton DF, Ford D, Peto J, Anderson K, Averill M, Stratton M, Ponder M, Pye C, Ponder BJA (1993): Genetic heterogeneity and localization of a familial breast-ovarian cancer gene on chromosome 17q12-q21. Am J Hum Genet 52:767-776.

Suthers GK, Wilson SR (1990): Genetic counseling in rare syndromes: a resampling method for determining an approximate confidence interval for gene location with linkage data from a single pedigree. Am J Hum Genet 47:53-61.

Weeks DE, Ott J (1989): Risk calculations under heterogeneity. Am J Hum Genet 45:819-821.